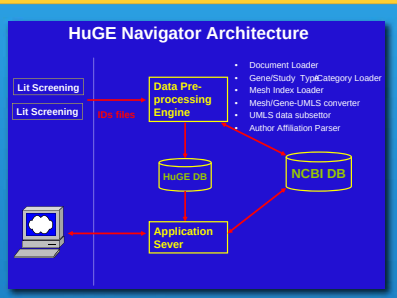
Implementing a UMLS-enabled Knowledgebase in Human Genome Epidemiology

W Yu, PhD, MS, A Yesupriya, M Gwinn, MD and MJ Khoury, MD, PhD
National Office of Public Health Genomics, CDC

Background

Human genome epidemiology (HuGE) is an evolving field of inquiry that uses systematic applications of epidemiologic methods and approaches in population based studies of the impact of human genetic variation on health and disease. In last the few years, the vast numbers of research findings in HuGE have been deposited in public domains such as PubMed. In 2000, the CDC National Office of Public Health Genomics (NOPHG) began an initiative to identify published literature in PubMed that is relevant to HuGE. To facilitate HuGE research, we have developed a knowledgebase system called HuGE Navigator, consisting of several different applications that use the Unified Medical Language System (UMLS) to maximize data interoperability and integration..



System architecture

The HuGE Navigator was built on three discrete modules that are loosely coupled. The data module contains all data in the database, the accessory utility module is responsible for a series of data transactions and manipulations, and the application module includes all applications in the system. To avoid versioning issues, we allow data entities from external data sources (e.g., UMLS Metathesaurus, Entrez Gene and MeSH tree), to be updated as needed without an overhaul of the entire system. Each application was built on top of this model, allowing for seamless navigation and easy plug-in of new applications.

Literature document indexing strategy

Document indexing is crucial for successful information retrieval. The accessory utility module downloads PubMed abstracts and their corresponding Medical Subject Heading (MeSH) terms and uploads them to the HuGE Navigator database automatically. The accessory utility module then maps the MeSH terms to UMLS concepts and indexes the document with the corresponding UMLS codes. The MeSH tree, a standardized hierarchical relationship between MeSH terms, has also been incorporated into the system to increase the sensitivity of information retrieval by including "children" terms. Gene information (symbol, name, aliases) from the NCBI Entrez Gene database is integrated into UMLS concept list to enhance the capacity of the UMLS Metathesaurus.

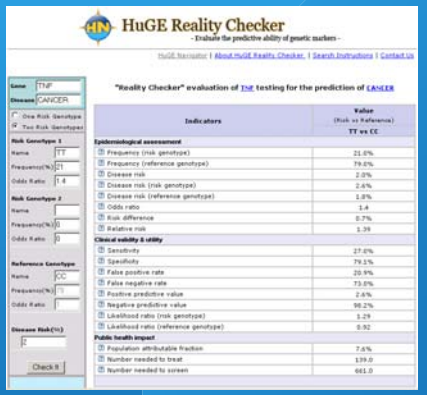
Benefits of UMLS implementation

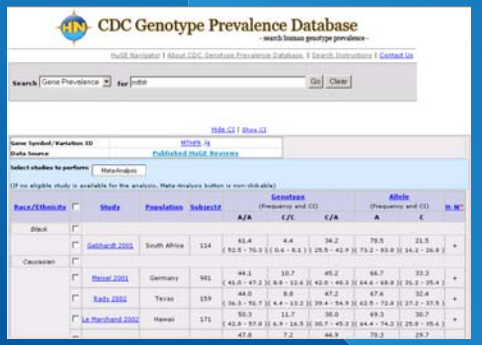
Because UMLS contains over 100 vocabularies from biomedical fields, many synonyms and variants of terms are collected in the Metathesaurus, which combined with UMLS indexing allows for robust free text searching. Incorporating synonyms into user queries increases the sensitivity of searching external databases (e.g. NCBI Gene Database, PubMed). Data interoperability is a big benefit of UMLS implementation.

http://wirm-sweb-srv3/HuGENavigator/startHuGENavigator.do

Improvement of performance by dynamic data subsetting

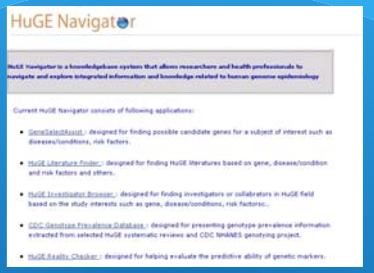
The 6 million unique concept names in UMLS could create performance issues if the table for the codes had to be queried directly. Even after removing non-English and retired concepts, the table still contains 3 million records. The multidisciplinary nature of human genome epidemiology precludes further subsetting by domain-specific criteria. To resolve this issue, we created an automatic UMLS concept subsetting process, populating the subset table dynamically whenever new MeSH terms were identified in the literature deposited in the database. The size of the UMLS subset data (23,000) was reduced dramatically, significantly improving performance.





User-friendly web interface

The HuGE Navigator interface was designed with a similar look and feel for each application facilitating easy navigation among applications and maximum information retrieval. The interface is designed to be simple and intuitive, so that the user knows how to perform each search without reading instructions.



Components of the system

- *GeneSelectAssist:* designed to help identify candidate genes for genetic epidemiology association studies.
- *HuGE Literature Finder:* designed for finding published literature on human genome epidemiology.
- *CDC Genotype Prevalence Database:* designed for presenting genotype prevalence information extracted from selected HuGE systematic reviews and the CDC NHANES genotyping project.
- *HuGE Investigator Browser:* designed for finding investigators or collaborators in human genome epidemiology
- *HuGE Reality Checker:* designed to help evaluate the predictive ability of genetic markers.

Future directions

The HuGE Navigator is an important new tool, supporting the global HuGE community's goal of research synthesis. UMLS is a good choice for indexing HuGE literature (see poster by Yesupriya, et al). Equipped with other UMLS sources such as Semantic Network and SPECIALIST Lexicon, we should be able to explore the published literature much more effectively and to infer more in-depth knowledge of human genome epidemiology.

Acknowledgement

We would like to express appreciation to Jim Arnzen for his technical support of the application server, and to Terrine Mathews for his support of the database.